MFMDQwen: Multilingual Financial Misinformation Detection Based on Large Language Model

Zhiwei Liu[1*] **Yuyan Wang**[1*] **Yuechen Jiang**[1] **Yupeng Cao**[2] **Tianlei Zhu**[3] **Xiaorui Guo**[4]
Zhiyang Deng[2] **Zhiyuan Yao**[2] **Xiao-Yang Liu**[3] **Jimin Huang**[1,5] **Sophia Ananiadou**[1,6]
[1]The University of Manchester [2]Stevens Institute of Technology [3]Columbia University
[4]The University of Edinburgh [5]The Fin AI [6]ELLIS Manchester
{zhiwei.liu,sophia.ananiadou}@manchester.ac.uk, ycao33@stevens.edu, tz2617@columbia.edu
{yuyan.wang-11,yuechen.jiang,jimin.huang}@postgrad.manchester.ac.uk
X.Guo-46@sms.ed.ac.uk, zdeng10@stevens.edu, zyao9@stevens.edu, xl2427@columbia.edu

Abstract

Financial misinformation poses significant threats to financial market stability and individuals' investment decisions. The multilingual environment and the inherent complexity of financial information present substantial challenges for Multilingual Financial Misinformation Detection (MFMD). Existing LLM-based approaches for financial misinformation detection primarily focus on English and a single financial misinformation detection task, which limits their ability to capture multilingual contexts and complex features. In this paper, we propose MFMDQwen, the first open-source LLM designed for MFMD tasks. Furthermore, we introduce MFMD4Instruction, the first instruction dataset supporting MFMD with LLMs, covering English, Chinese, Greek, and Bengali. We also construct MFMDBench, a benchmark dataset for evaluating the MFMD capabilities of LLMs. Experimental results on MFMDBench demonstrate that our model outperforms existing open-source LLMs. The project is available at https://github.com/lzw108/FMD.

1 Introduction

In the financial domain, the accuracy of information is crucial for investment decisions, financial decision-making, and the stability of financial markets, especially in multilingual environments (Rangapur et al., 2023b; Liu et al., 2026a). However, the rapid growth of digital media has exacerbated the spread of financial misinformation. Such misinformation, including deceptive investment advice and misleading statements that distort financial markets, can influence asset prices and broader economic sentiment, thereby posing significant risks (Nag, 2025). The complex characteristics of financial information, such as minor numerical changes, amplified sentiment, and reversed causality, present substantial challenges for automated detection (Jiang et al., 2026). Time-consuming manual inspection is clearly unsuitable for detecting the rapidly evolving and large volume of financial misinformation, and the development of large language models (LLMs) has made automated detection methods possible.

In recent years, LLMs with massive numbers of parameters have emerged as a novel approach to tackling various problems in the financial domain, achieving remarkable results (Li et al., 2023). This includes applications in financial misinformation detection, such as FMDLlama (Liu et al., 2025d), and the creation of financial misinformation datasets like FinFact (Rangapur et al., 2023a), FinDver (Zhao et al., 2024), and RFCBench (Jiang et al., 2026), as well as studies on bias (Liu et al., 2026a). Nevertheless, most research still focuses on English data, overlooking the more complex and higher-risk problem of multilingual financial misinformation.

To tackle these issues, we construct the first instructing-tuning dataset for multilingual financial misinformation detection (MFMD4Instruction) to support LLMs' supervised fine-tuning (SFT). Based on MFMD4Instruction, we developed the first open-source multilingual financial misinformation model through SFT to support the MFMD task across multiple languages, including English, Chinese, Greek, and Bengali. To evaluate the financial misinformation verification ability of LLMs, we also built a benchmark for the detection of multilingual financial misinformation (MFMDBench). We evaluated MFMDQwen and numerous baselines on MFMDBench, and the results demonstrate that MFMDQwen achieved the best overall performance across all nine datasets.

Our main contributions are as follows:

(1) We construct MFMD4Instruction, the first multilingual financial misinformation dataset for SFT of LLMs.

(2) We develop MFMDQwen, the first open-source LLM for multilingual financial misinfor-

[*]These authors contributed equally to this work.

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 75–82

mation detection.

(3) We build MFMDBench, the first benchmark to evaluate the verification ability of multilingual financial misinformation of LLMs, including 9 tasks and covering 4 languages. The results on MFMDBench demonstrate that our model outperforms other open-source LLMs.

2 Related work

2.1 Financial Misinformation Detection

In finance, where accurate information is essential for market stability, and trust, digital media has accelerated the spread of misinformation (Rangapur et al., 2023b). Recent studies have explored automated financial misinformation detection with LLMs. Rangapur et al. (2023a) propose the dataset for financial fact checking and explanation generation, and evaluate the ability of several LLMs. FMDLlama applies instruction tuning to adapt LLMs to this task (Liu et al., 2025d), while the FinFact workshop brought together diverse approaches on a shared dataset (Liu et al., 2025c). Other work addresses financial data scarcity through general-domain augmentation, evidence generation, and few-shot retrieval (Lee and Park, 2025), or extends detection to multimodal settings by combining text with image-derived descriptions (Luo et al., 2025). Beyond detection, Cao et al. (2025) improve financial reasoning by combining retrieved evidence with a Financial Chain-of-Thought framework, and FinDVer provides a benchmark for explainable claim verification on long, hybrid financial documents, showing that even GPT-4o trails human experts (Zhao et al., 2024). More recently, MFMDBench (Liu et al., 2026a) introduced a multilingual benchmark for financial misinformation detection across diverse cultural contexts. RFC-BENCH (Jiang et al., 2026) further investigates professional financial misinformation detection in a reference-free setting. Both benchmarks reveal that current LLMs still exhibit substantial limitations when handling financial misinformation detection in complex and professional scenarios.

2.2 Open-sourced Large Language Models

LLMs have been widely applied across various fields and have achieved good results. For example, the ChatGPT series (OpenAI, 2025), Deepseek (Liu et al., 2025a), and LLama (Dubey et al., 2024). There is also a lot of research dedicated to open-sourcing domain-specific LLMs, such as LLMs for finance (Wu et al., 2023; Xie et al., 2023), health (Chen et al., 2024; Xiao et al., 2025), misinformation detection (Liu et al., 2025b, 2026b), and so on. These open-source models have facilitated in-depth research in their respective domains. For financial misinformation, the existing FMDllama (Liu et al., 2025d) only supports English and cannot be applied to multilingual financial misinformation detection. Therefore, this paper develops MFMDQwen for multilingual misinformation detection.

3 Methods

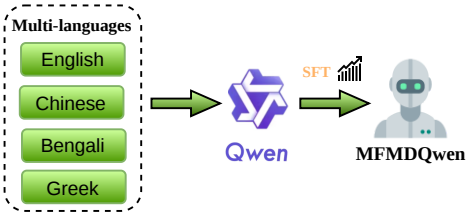

Figure 1: The architecture of MFMDQwen.

3.1 Task formalization

We formulate financial misinformation detection as a generative task, leveraging a generative model as the foundation. Specifically, we adopt an autoregressive language model $\mathbf{P}_\phi(y \mid x)$ parameterized by pre-trained weights ϕ. This model is capable of simultaneously handling multiple multilingual financial misinformation detection tasks. Each task t is defined as a set of context-target pairs:

$$D_t = \{(q_i^t, r_i^t)\}_{i=1}^{N_t},$$

where: q_i^t is a context token sequence that includes the task description, input text, and query; r_i^t is a target token sequence that represents the final answer.

To enable multi-task learning, all task datasets are merged into a single dataset. The model is then trained to maximize the conditional likelihood of the target sequences given the contexts:

$$\max_{\phi} \sum_{t} \sum_{i=1}^{N_t} \log \mathbf{P}_\phi(r_i^t \mid q_i^t).$$

Through this setup, the model learns to generate final answers, improving prediction accuracy via SFT.

3.2 MFMD4Instruction

3.2.1 Raw Data Collection

Based on MFMDScen (Liu et al., 2026a) and RFC (Jiang et al., 2026), we compile a multilingual financial misinformation corpus from nine existing sources spanning four languages: English, Chinese, Bengali, and Greek. These datasets cover a range of related tasks, including claim verification, fact-checking, misinformation detection, and social media manipulation detection.

English. **FinDVer** (Zhao et al., 2024) contains 700 financial claim verification instances derived from financial reports. Each instance consists of a financial statement paired with relevant evidence paragraphs from the corresponding report. The dataset is balanced, with 350 *entailed* and 350 *refuted* samples. **RFC-BENCH** (Jiang et al., 2026) is a paragraph-level benchmark for reference-free counterfactual financial misinformation detection. It contains 1,826 original–perturbed paragraph pairs derived from real Yahoo Finance articles covering 223 U.S. stocks. Perturbations are generated via GPT-4.1 under four manipulation types (Directional Flipping, Numerical Perturbation, Sentiment Amplification, and Causal Distortion) and validated through multi-stage expert review and dual-annotator evaluation to ensure label reliability.

Chinese. **CHEF** (Hu et al., 2022) is a Chinese fact-checking dataset with 1,188 samples, each comprising a claim and associated evidence texts. Its labels are *supported*, *refuted*, and *not enough information*. **MDFEND** (Nan et al., 2021) contains 1,321 Weibo posts labeled as either *real* (959 samples) or *fake* (362 samples).

Bengali. **Bengali Manipulation** (Kamruzzaman et al., 2023) (**BanMANI**) includes 101 samples for manipulated social media detection. Each sample pairs an original news article with a related social media post and is labeled as *MANI* (52 samples) or *NO_MANI* (49 samples).

Multilingual. **Global4Languages** (Liu et al., 2026a) provides aligned claim verification samples in English, Chinese, Bengali, and Greek, with 144 instances per language. Each sample contains a claim and its scenario context, labeled as *true* (23 samples) or *false* (121 samples).

Dataset	Language	MFDM4Instruction	MFMDBench	Total
FinDVer	EN	554	140	694
CHEF	ZH	894	238	1132
MDFEND	ZH	1044	265	1309
Bengali	BN	80	21	101
Global-EN	EN	115	29	144
Global-ZH	ZH	115	29	144
Global-BN	BN	115	29	144
Global-GR	GR	115	29	144
RFC	EN	1805	1652	3457
Total	-	4837	2432	7269

Table 1: Statistics of the datasets. EN: English, ZH: Chinese, BN: Bengali, GR: Greek.

3.2.2 Construction of Base Data for Complex Reasoning

To prepare the raw datasets for complex reasoning path construction, we convert each dataset into a unified instruction-tuning format. Each sample is transformed into a structured record containing: (1) a **task description** specifying the detection task and expected output format, (2) the **input fields** consisting of the claim or content together with any supporting evidence or contextual information. All templates can be found at Appendix A.

3.3 MFMDQwen

Figure 1 presents the overview of MFMD-R. We first construct the reasoning paths (MFMD4Instruction). We then built MFMD-R based on Qwen-3-8B (Yang et al., 2025) using the MFMD4Instruction dataset. The model is trained in two stages: SFT followed by RL. During SFT, the learning rate is set to 1×10^{-5} with a warmup ratio of 0.1. Training runs for 2 epochs with a batch size of 128. DeepSpeed ZeRO-3 optimization is used with CPU parameter offloading to reduce GPU memory usage. The maximum input sequence length is 24k tokens and the maximum output length is 8k tokens. Full-parameter training is conducted on four NVIDIA L40s GPUs (48 GB).

4 Experiments

4.1 Baseline models

Our extensive evaluation of open-source and proprietary LLMs included the following reasoning-focused models: Qwen3 reasoning variants (8B-R, 14B-R, and 32B-R), Qwen3 no-reasoning models (8B, 14B, and 32B) (Yang et al., 2025), Qwen2.5-72B-Instruct (Qwen et al., 2025), Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct (Dubey et al., 2024). We also compare domain-specific financial misinformation LLMs (i.e. FMDLlama (Liu et al., 2025d))

Models	GlobalGr		GlobalBe		GlobalCh		GlobalEn		CHEF		MDFEND		Bengali		FinDver		RFC		Ave.
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	F1
Qwen3-8b-R	0.862	0.710	0.655	0.396	**0.862**	**0.710**	0.793	0.638	0.420	0.207	0.668	0.652	0.905	0.905	0.807	0.806	0.668	0.652	0.631
Qwen3-14b-R	0.897	0.756	0.759	0.607	0.862	0.628	**0.897**	**0.756**	0.429	0.272	0.555	0.553	0.762	0.760	0.814	0.814	0.555	0.553	0.633
Qwen3-32b-R	0.793	0.638	0.793	0.685	**0.862**	**0.710**	0.828	0.671	0.424	0.275	0.653	0.639	0.952	0.952	**0.857**	**0.857**	0.653	0.639	0.674
Qwen3-8b	0.862	0.710	0.828	0.671	0.759	0.431	0.862	0.628	0.462	0.320	0.608	0.604	0.952	0.952	0.750	0.750	0.608	0.604	0.630
Qwen3-14b	0.759	0.431	0.793	0.685	0.828	0.671	0.828	0.671	0.429	0.272	0.623	0.617	0.762	0.741	0.793	0.793	0.623	0.617	0.611
Qwen3-32b	0.862	0.628	0.828	0.671	0.828	0.671	0.759	0.607	0.458	0.310	0.709	0.689	0.952	0.952	0.793	0.793	0.709	0.689	0.668
Qwen2.5-72b	0.862	0.710	0.793	0.565	0.759	0.540	0.828	0.671	0.521	0.320	0.857	0.819	0.952	0.952	0.843	0.841	0.551	0.543	0.662
Llama3.1-8b	0.621	0.466	0.207	0.133	0.690	0.427	0.759	0.463	0.500	0.203	0.664	0.429	0.429	0.276	0.721	0.484	0.664	0.429	0.368
Llama3.3-70b	**0.931**	**0.879**	0.759	0.470	0.759	0.653	0.793	0.638	0.471	0.279	0.815	0.763	**1.000**	**1.000**	0.821	0.818	0.529	0.466	0.663
FMDLlama	0.759	0.431	0.069	0.051	0.828	0.594	0.793	0.383	0.580	0.193	0.272	0.143	0.429	0.299	0.614	0.402	0.272	0.143	0.293
MFMDQwen	0.862	0.758	**0.897**	**0.803**	**0.862**	**0.710**	**0.897**	**0.756**	**0.849**	**0.843**	**0.932**	**0.913**	0.952	0.952	0.700	0.673	**0.954**	**0.954**	**0.818**

Table 2: Results on MFMDBench.

4.2 Evaluation methods

We uses metrics such as Accuracy, Macro-F1 for misinformation detection evaluation.

4.3 Results

Table 2 presents the experimental results on the MFMDBench dataset. In the following analysis, we primarily focus on the F1 score. As shown in the table, MFMDQwen attains top scores on 6 of 9 benchmarks, including models of comparable size as well as larger LLMs with 14B, 32B, and 72B parameters. These results indicate that MFMDQwen demonstrates clear advantages in multilingual tasks and validate the effectiveness of SFT training. This strategy significantly enhances the multilingual representation capabilities of LLMs. Additional confusion matrix visualizations are provided in Appendix B.

It is also worth noting that some LLMs occasionally produce uncertain or irrelevant responses when answering questions, which partially explains the relatively lower F1 scores. This phenomenon is particularly evident in Llama3.1-8B and FMDLlama, which frequently generate responses indicating that they are unable to provide an answer. This behavior may be related to their relatively strict safety or protection mechanisms. Additionally, models with reasoning capabilities do not always outperform non-reasoning models of the same size, as the additional reasoning steps may lead to overthinking and ultimately produce incorrect answers.

Furthermore, performance differences across languages reveal additional insights into model behavior. MFMDQwen shows the most consistent gains on Chinese datasets (CHEF and MDFEND), where it significantly outperforms all baselines. This suggests that the proposed training strategy is particularly effective in handling noisy, context-dependent misinformation commonly found in social media environments. In contrast, on English datasets, the performance gap between MFMDQwen and strong LLM baselines is relatively smaller. For instance, on FinDVer, some larger models achieve comparable or even better results, indicating that structured evidence-based judgment remains a strength of general-purpose LLMs.

For low-resource languages such as Bengali, most models achieve near-saturated performance, which can be attributed to the limited dataset size and reduced task complexity. On the GlobalGr dataset, MFMDQwen ranks second, slightly below Llama3.3-70B, while still outperforming most other baselines. Notably, it achieves the best performance on GlobalBe. These results further validate the effectiveness of the proposed fine-tuning strategy, showing that it enables strong cross-lingual generalization and competitive performance even against substantially larger models.

Overall, MFMDQwen demonstrates superior cross-lingual stability compared to both general-purpose LLMs and the domain-specific FMDLlama model. While FMDLlama is tailored for financial misinformation detection, its performance varies significantly across languages, indicating limited multilingual generalization. In contrast, MFMDQwen benefits from supervised fine-tuning on multilingual datasets, enabling it to capture more robust and language-invariant misinformation patterns across diverse linguistic settings.

5 Conclusion

In this paper, we propose MFMDQwen, the first open-sourced LLM for Multilingual Financial Misinformation Detection (MFMD). We also construct a multi-task multilingual financial misinformation dataset (MFMD4Instruction) and an MFMD evaluation benchmark (MFMDBench). We conduct a comprehensive evaluation of MFMDQwen and a variety of LLMs on the MFMDBench benchmark. The results show that MFMDQwen achieves the

best performance on 6 out of 9 datasets, demonstrating its strong capability in MFMD tasks.

In future work, we plan to expand MFMD4Instruction and MFMDBench to include more languages, integrate additional financial-related tasks, and further improve the overall performance of MFMDQwen.

Limitations

Due to restricted computational resources and cost, we only carried out instruction-tuning/evaluation of multilingual financial misinformation detection tasks using 8b/14b/32b LLMs. As such, we have not considered the impact of using larger models on the MFMDBench tasks.

Acknowledgments

This research was supported by the NVIDIA Academic Grant Program using 32K A100 GPU-hours on Brev. The authors acknowledge The Fin AI community for its research support, feedback, and collaborative environment that contributed to this work.

References

Yupeng Cao, Haohang Li, Yangyang Yu, and Shashidhar Reddy Javaji. 2025. Capybara at the financial misinformation detection challenge task: chain-of-thought enhanced financial misinformation detection. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 321–325.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. 2022. Chef: A pilot chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376.

Yuechen Jiang, Zhiwei Liu, Yupeng Cao, Yueru He, Ziyang Xu, Chen Xu, Zhiyang Deng, Prayag Tiwari, Xi Chen, Alejandro Lopez-Lira, et al. 2026. All that glisters is not gold: A benchmark for reference-free counterfactual financial misinformation detection. *arXiv preprint arXiv:2601.04160*.

Mahammed Kamruzzaman, Md Minul Islam Shovon, and Gene Kim. 2023. Banmani: A dataset to identify manipulated social media news in bangla. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 51–58.

Dongjun Lee and Heesoo Park. 2025. Dunamu ml at the financial misinformation detection challenge task: improving supervised fine-tuning with llm-based data augmentation. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 297–301.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Zhiwei Liu, Yupen Cao, Yuechen Jiang, Mohsinul Kabir, Polydoros Giannouris, Chen Xu, Ziyang Xu, Tianlei Zhu, Tariquzzaman Faisal, Triantafillos Papadopoulos, et al. 2026a. Same claim, different judgment: Benchmarking scenario-induced bias in multilingual financial misinformation detection. *arXiv preprint arXiv:2601.05403*.

Zhiwei Liu, Runteng Guo, Baojie Qu, Yuechen Jiang, Min Peng, Qianqian Xie, and Sophia Ananiadou. 2026b. Raar: Retrieval augmented agentic reasoning for cross-domain misinformation detection. *arXiv preprint arXiv:2601.04853*.

Zhiwei Liu, Paul Thompson, Jiaqi Rong, and Sophia Ananiadou. 2025b. Conspemollm-v2: A robust and stable model to detect sentiment-transformed conspiracy theories. *arXiv preprint arXiv:2505.14917*.

Zhiwei Liu, Keyi Wang, Zhuo Bao, Xin Zhang, Jiping Dong, Kailai Yang, Mohsinul Kabir, Polydoros Giannouris, Rui Xing, Park Seongchan, et al. 2025c. Finnlp-fnp-llmfinlegal-2025 shared task: financial misinformation detection challenge task. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 271–276.

Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2025d. Fmdllama: Financial misinformation detection based on large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1153–1157.

Zheyang Luo, Guangbin Zhang, Jiahao Xiao, Xuankang Zhang, Yulin Dou, and Jiangming Liu. 2025. Fmdmllama at the financial misinformation detection challenge task: multimodal reasoning and evidence generation. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 277–282.

Shakalya Nag. 2025. Financial misinformation and trading manipulation using large language models (llms). In *International IOT, Electronics and Mechatronics Conference*, pages 1–15. Springer.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3343–3347.

OpenAI. 2025. Introducing gpt-4.1. Official OpenAI announcement for the GPT-4.1 model family.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Aman Rangapur, Haoran Wang, and Kai Shu. 2023a. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *Preprint*, arXiv:2309.08793.

Aman Rangapur, Haoran Wang, and Kai Shu. 2023b. Investigating online financial misinformation and its consequences: A computational perspective. *arXiv preprint arXiv:2309.12363*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Mengxi Xiao, Kailai Yang, Pengde Zhao, Enze Zhang, Ziyan Kuang, Zhiwei Liu, Weiguang Han, Shu Liao, Lianting Huang, Jinpeng Hu, et al. 2025. Mentrasuite: Post-training large language models for mental health reasoning and assessment. *arXiv preprint arXiv:2512.09636*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024. Finver: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752.

A Prompt templates

Based on MFMDScen (Liu et al., 2026a) and RFC (Jiang et al., 2026), we define the following prompts as templates:

Prompt template for Global4Languages:

Task Description: Determine whether the claim is 'True' or 'False'.
Claim: *[Claim]*

Prompt template for FinDVer:

Task Description: Assess the truthfulness of the given statement by determining whether it is entailed or refuted based on the provided financial document. Output the entailment label ('entailed' or 'refuted') of the claim.
Claim: *[Claim]*
Relevant Financial Report: *[Document]*

Prompt template for MDFEND:

Task Description: Determine whether the following content is 'real' or 'false'.
Content: *[Content]*

Prompt template for CHEF:

Task Description: Label each claim based on the evidence provided. Choose one of the following three labels: Supported, which means there is sufficient evidence showing the claim is supported; Refuted, which means there is sufficient evidence showing the claim is refuted; Not enough information, which means the evidence is insufficient to determine whether the claim is supported or refuted.
Claim: *[Claim]*
Evidence: *[Evidence]*

Prompt template for BanMANI:

Task Description: Determine whether the social media post is manipulated or not manipulated based on the original news. Output 'MANI' in case the post is manipulated from the original news article, or output 'NO_MANI' otherwise.
Original News: *[Original News]*
Social Media Post: *[Social Media Post]*

Prompt template for RFC:

Task Description: You are a financial misinformation detector. Please check whether the following information is true or false and output the answer [true/false].
News: *[News]*

B Confusion Matrix Visualizations

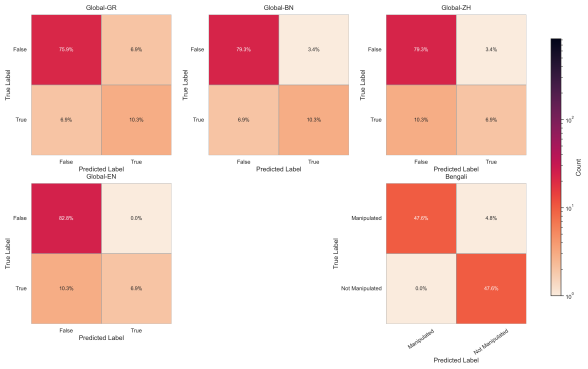

Figure 2: Confusion matrices for five binary classification datasets (GlobalGr, GlobalBe, GlobalCh, GlobalEn, and Bengali). We apply logarithmic normalization to improve visual contrast across datasets with different scales. Rows correspond to ground-truth labels and columns correspond to model predictions.

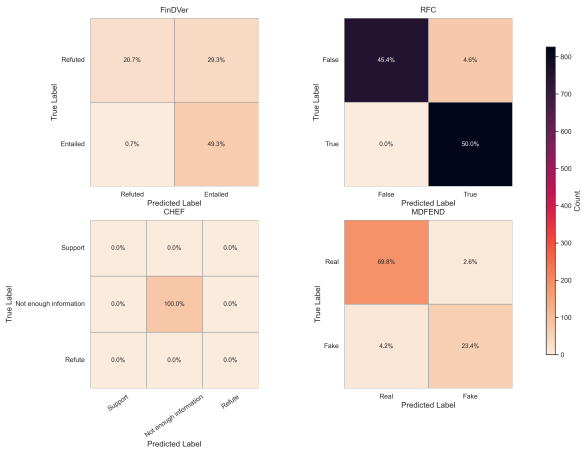

Figure 3: Confusion matrices for four datasets with heterogeneous label spaces (Chef, MDFEND, FindVer, and RFC). A shared linear normalization is used to preserve the relative magnitude of counts across datasets.

C Error Case Study

We manually inspected representative prediction errors across datasets and identified five recurring error types. These errors suggest that the model does not fail uniformly; rather, its mistakes stem from distinct weaknesses in polarity control, numerical reasoning, evidence grounding, and fine-grained semantic comparison.

C.1 Error Type 1: Label Polarity Reversal and Prior-Driven Guessing

A common failure mode in multilingual true/false datasets is label polarity reversal, in which the model predicts the opposite label despite the dataset's simple binary format. In several cases, the model appears to rely on surface plausibility or world-knowledge priors rather than on the benchmark label.

Case Study. In **Global-EN**, the model predicts *False* for the claim: *"Amazon solicited donations from the public to pay sick leave to contractors and seasonal workers during the COVID-19 pandemic."* The gold label is *True*. Similar polarity reversals also appear in **Global-GR**, **Global-BN**, and **Global-ZH** for translated variants of the same claim. This cross-lingual consistency suggests that the error is not solely due to translation difficulty but to a systematic tendency to default to a plausible-sounding negative judgment.

C.2 Error Type 2: Numerical Reasoning and Arithmetic Errors

The model frequently fails on claims that require explicit calculation, especially percentage change, percentage decrease, or net value derived from financial figures. These errors indicate weak arithmetic grounding even when all necessary numbers are explicitly present in the document.

Case Study. In **FinDVer**, the model predicts *Entailed* for the claim: *"The percentage increase in the working capital deficit from March 31, 2023, to December 31, 2023, is approximately 38.33%."* The gold label is *Refuted*. The report states that the working capital deficit increased from $84,255 to $116,603. This requires an explicit percentage-change calculation, and the model incorrectly accepts the claimed value rather than verifying it against the numbers.

C.3 Error Type 3: Table Parsing and Financial Statement Grounding Errors

Another major source of errors is incorrect grounding in semi-structured financial tables. The model often extracts the wrong row, wrong column, or wrong derived quantity, especially when multiple time periods and measures are presented together.

Case Study. In **FinDVer**, the model predicts *Entailed* for the claim: *"The total operating margin percentage for Kennametal over the six months ended December 31, 2023, is 7.4%."* The gold label is *Refuted*. The evidence is presented in a multi-column table with total sales and total operating income for different periods. Correct judgment requires selecting the correct six-month values and computing the margin from them. The model appears to accept the stated percentage without robustly grounding it in the table.

C.4 Error Type 4: Partial Evidence Matching and Missing Critical Detail

The model also fails when a claim is partially supported by the evidence but includes one critical, incorrect detail. In such cases, it appears to match the generally relevant topic while overlooking the exact attribute that determines the label.

Case Study. In **FinDVer**, the model predicts *Entailed* for the claim: *"In 2023, AMN Healthcare saw 54% of its consolidated revenue flow through Managed Services Programs, while Kaiser Foundation Hospitals accounted for approximately 27% of its total consolidated revenue."* The gold label is *Refuted*. The first part of the claim is supported by the report, but the second part is not: the report states that Kaiser accounted for approximately **17%**, not 27%. The model likely overrelied on partial support from the first clause and failed to carefully verify the second clause.

C.5 Error Type 5: Manipulation Detection Failure under Subtle Lexical Mismatch

For manipulation detection, the model struggles when a social media post preserves the original news's general topic but alters a key factual detail. These errors show weak sensitivity to subtle semantic distortion rather than complete topic mismatch.

Case Study. In **Bengali**, the model predicts *Not Manipulated* for the post: *"City Bank-American Express launches the first airport lounge."* The gold label is *Manipulated*. The original news states that the newly launched lounge is their **second international lounge**, while also mentioning that the company had launched the country's first airport lounge earlier. The manipulated post changes the event-specific fact from *second* to *first*. The model appears to match the overall topic correctly but misses the crucial ordinal inconsistency.

C.6 Discussion

Overall, the observed errors indicate that the model's weaknesses are not limited to a single dataset or language. Instead, they cluster around a small set of recurring reasoning failures: polarity instability in binary classification, weak arithmetic verification, fragile grounding in financial tables, over-acceptance under partial evidence overlap, and insufficient sensitivity to subtle factual distortions. These findings suggest that future improvements should focus not only on general instruction following but also on explicit verification mechanisms for numbers, table structure, and claim-level fact alignment.